B2B pricing

Force Edouard

Approche bayésienne de la régression logistique

9 Avril 2021

Situation

L'article ici nous présente une méthode pour lier l'approche bayésienne à la régression logistique. La problématique est la suivante :
Un vendeur possède un nombre n d'objets à vendre. Le vendeur vend les articles à un certain prix $(p_k)_{k\in[|1,n|]}$. En face, un acheteur veut acheter un ou plusieurs produits. L'acheteur peut alors proposer deux réponses aux produits: soit il achète, soit il refuse.

Variables

On note alors comme suivante les variables :

- $Y \in \{0,1\}^n$
- $X \in \mathbb{R}^{n*m}$
- $p \in \mathbb{R}_+^n$

Y correspond à la réponse de l'acheteur, $Y = 1$ correspond à une réponse positive de l'acheteur. X est notre dataset. X correspond en faite aux données sur la vente.

Dataset X

X est en fait la connaissance des informations que possède le vendeur sur une multitude de caractéristiques. Dans un premier temps, certaines features de X dépendent fortement du prix p. Ensuite, X a une dépendance suivant le produit, des informations sur le vendeur, etc.

On peut voir notre problème selon plusieurs cas. On peut supposer que X est connu par le vendeur, et donc on peut trouver le lien entre X et Y. On peut également supposer que X n'est pas connu en entier, avec des valeurs manquantes, c'est à dire que le vendeur ne connait pas toutes les infos sur par exemple, l'acheteur, l'envie des acheteurs, etc.

Modèle

On suppose que X est connu. On va utiliser la régression logistique pour décrire nos données. On pose la probabilité de vendre $\mathbb{P}(Y = 1)$ comme une fonction:

$$\rho(x, \beta) = \frac{1}{1 + e^{-\beta^T x}}$$

où β est un vecteur qui est inconnu. Ce vecteur permettra de définir:

$$\log \frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$$

Revenu

Le but du vendeur n'est pas uniquement que tous les acheteurs donnent une réponse positive. Le but étant d'avoir le plus de revenu selon le meilleur prix p de vente. On s'intéresse alors à la fonction revenue suivante :

$$R(p, x, \beta) = p \,.\, \rho(x, \beta)$$

$p^\star = argmax_p R(p, x, \beta)$ représente le prix optimal pour le revenu du vendeur. x ici dépend également de p.

Modèle Bayésien : Prior

Notre recherche ici est portée sur l'inconnu β. Pour approcher beta, on va alors utiliser une approche bayésienne. On va poser un prior sur β:

$$\beta \sim \mathcal{N}(\theta, \Sigma).$$

β suit une loi normale multidimensionnelle de moyenne θ et de covariance Σ. On choisit cette loi pour une raison principale : les liens de correlations entre les variables sont bien représentés dans la matrice Σ

Modèle Bayésien : Posterior

Dans la régression logistique, on note la probabilité sur Y par rapport à x et β comme l'expression suivante :

$$\mathbb{P}_x(Y|\beta) = \frac{1}{1 + e^{-sign(Y)\beta^T x}} = \ell(H(\beta, x))$$

où $\ell(z) = \frac{1}{1+e^{-z}}$ et $H(\beta, x) = -sign(Y)\beta^T x$

Modèle Bayésien : Posterior

Le posterior peut alors s'écrire comme :

$$\pi(\beta|Y,x) \propto \frac{1}{1+e^{-sign(Y)\beta^T x}}|\Sigma|^{-1/2}e^{-\frac{1}{2}(\beta-\theta)^T\Sigma^{-1}(\beta-\theta)}$$

Cette forme n'est clairement pas le terme général d'une loi usuelle. Il est donc difficile de simuler β suivant cette expression. On va alors l'approcher à l'aide de la divergence de Kullback-Leibler. Pour garder les correlations "visiblent", on veut essayer d'approcher $\pi(\beta|Y,x)$ par une loi normale multidimensionnelle.

Approche Bayésienne : Divergence KL

On introduit alors une densité d'une loi normale Q de paramètre (θ', Σ') et on veut alors trouver les paramètres optimaux $(\theta'^{\star}, \Sigma'^{\star})$ qui minimise :

$$\mathcal{D}^{KL}(Q||P) = \mathbb{E}_Q\left(\log \frac{Q(\beta)}{P_x(\beta)}\right)$$

Gradient KL

Pour trouver le minimum de ceci, on commence par dériver la Divergence de Kullback. Par proposition, on obtient :

$$\nabla \mathcal{D}^{KL}(Q||P) = \mathbb{E}_Q \left[\log(1 + e^{-H(\beta, Y)}) \right] + \nabla \mathcal{D}^{KL}(Q||P_0)$$

où P_0 est la densité de la loi $\mathcal{N}(\theta, \Sigma)$. On pourrait utiliser des méthodes de descentes de gradient connu pour trouver notre minimum. Mais d'un point de vue computationnelle, cela peut prendre du temps, puisque la compléxité est de l'ordre de $m^2 \times m$, m étant le nombre de features.

Optimisation du gradient KL

L'article nous propose alors une autre approche, à l'aide des formules de Sherman-Morrison-Woodbury. On pose :

$$\theta' = \theta + \frac{v(Y - \frac{1}{2}) - x^T\theta)}{v + x^T\Sigma x}\Sigma x$$
$$\Sigma' = \Sigma - \frac{\Sigma x x^T \Sigma}{v + x^T\Sigma x}$$

où v est un paramètre réel. L'optimisation revient alors à chercher v qui minimise la divergence de Kullback.

Descent de Gradient

L'article nous propose alors d'utiliser la méthode d'approximation stochastique de Robbins-Monro qui approche le gradient selon v de la divergence de Kullback :

$$\nabla_v \mathcal{D}^{KL}(Q||P) = \left(\frac{v(Y-\frac{1}{2}) - x^T\theta)}{v + x^T\Sigma x}\right)\frac{(Y-\frac{1}{2})x^T\Sigma x + x^T\theta)}{(v+x^T\Sigma x)^2}(x^T\Sigma x)^2 - \frac{x^T\Sigma x}{2v}\frac{tr(xx^T\Sigma)}{(v+x^T\Sigma x)^2} + \hat{G}$$

où $\hat{G}$ est l'approximation stochastique de $\nabla_v\left(x^T\theta' + \sqrt{x^T\Sigma' x}Z(\omega)\right)$, avec $Z(\omega)$ une réalisation d'une loi normal centrée réduite. On réalise alors une descente de gradient classique pour obtenir une approximation de $v^\star$.

Implémentation à partir d'un rang n

On va supposer ici dans notre travail que la loi de β a pour paramètre (θ^n, Σ^n). On connait alors ici les n premières observations, c'est à dire les n premières réponses Y des acheteurs, et le dataset X connu jusqu'au rang n. Puis on observe alors x^{n+1} et Y^{n+1} et on applique alors ce que l'on a vu au dessus pour trouver (θ', Σ'). On peut regarder ici le code sur Python.

Jeu de validation et test

On remarque ici que en passant par l'approche bayésienne, le test de validation nous donne une précision de 100%, mais que le jeu de test nous donne une moins bonne précision (75%) que la version LogisticRegression de sklearn (82%).

LR et l'approche bayésienne

- La première différence se fait du point de vue computationnelle. En effet, la régression logistique de la bibliothèque sklearn est une méthode directe, brut et qui n'est pas adapté à la mise à jour des données. En effet, pour chaque nouvelle donnée, il faut réajuster le modèle avec toutes les données.
- A l'inverse, la méthode vue au dessus et sur Python est adaptative. On peut continuer à donner de nouvelles données et mettre à jour directement notre modèle.

LR et l'approche bayésienne

- On peut s'interroger de l'approximation du posterior par une loi normal multidimensionnelle. Est-elle justifiée ? Il faudrait regarder plus en détail les formules de Sherman-Morrison-Woodbury sur la mise à jour des θ' et Σ'. Bien que notre estimation du gradient soit sans bias, nous ne connaissons pas la disparité de celle-ci.
- Les proportions du nombre de réponses positives et négatives des acheteurs sont également à prendre en compte.
- La dépendance des variables suivantes le prix p. Ici dans le dataset choisit, la dépendance envers une constante existe mais est faible, comme on a pu le voir dans les figures.

Précision LR / KL

- La précision est également différente suivant les différentes méthodes.

Table 1 Performance metrics from three statistical models on the training and test data.

Metrics	Training Data			Test Data		
	LR	KL	VB	LR	KL	VB
Accuracy	0.867	0.863	0.858	0.828	0.825	0.823
AUC	0.871	0.858	0.842	0.851	0.839	0.827
F1 Score	0.445	0.441	0.471	0.439	0.439	0.479
Precision	0.643	0.616	0.566	0.643	0.616	0.591
Recall	0.340	0.343	0.403	0.333	0.341	0.403

La 'precision', c'est à dire le taux de vrai positif est meilleur pour la LR que pour l'approche bayésienne. Le vendeur préfère savoir combien il est sûr de vendre plutôt que de savoir si il va vendre ou non.